DOCUMENT RESUME

ED 312 309 TM 014 106

AUTHOR Marso, Ronald N.; Pigge, Fred L.
TITLE Staff Development Implications from a State-Wide Assessment of Classroom Teachers' Testing Skills and Practices.
PUB DATE Oct 89
NOTE 19p.; Paper presented at the Annual Meeting of the Mid-Western Educational Research Association (Chicago, IL, October 19-21, 1989).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Administrators; Classroom Techniques; Educational Assessment; *Elementary School Teachers; Elementary Secondary Education; Principals; *Secondary School Teachers; *Staff Development; State Surveys; *Supervisors; Teacher Effectiveness; *Teacher Made Tests; Teacher Supervision; Teaching Methods; *Test Construction; Test Use
IDENTIFIERS Ohio; *Teacher Competencies

ABSTRACT

This paper presents staff development implications for elementary and secondary school teachers which were gleaned from assessments of classroom teachers' testing needs, resources, practices, and proficiencies made by 586 supervisors and principals and 326 teachers in an Ohio study. These assessments included direct analyses of teacher-made tests as well as perceptual assessments of teachers' testing needs and proficiencies. It was generally agreed that testing proficiencies were not adequate to meet classroom needs. Analyses revealed that typical teachers gave 50 or more formal teacher-made tests each year, for which they wrote most of their own questions. Matching exercises on teacher-made tests were particularly prone to error. Most teacher-made tests, except in mathematics and science, functioned primarily at the knowledge level. Administrators' and teachers' perceptual assessments of teachers' testing skills were negatively correlated with the results of direct analysis of teacher testing skills as displayed in their teacher-made tests. Standardized tests were reported to be less needed in the secondary than in the elementary grades. Schools had very limited support services, such as typing and duplication, available to support the testing responsibilities of their teachers. Two tables present study data. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Staff Development Implications From A State-Wide Assessment of Classroom Teachers' Testing Skills and Practices

Ronald N. Marso and Fred L. Pigge
College of Education and Allied Professions
Bowling Green State University
Bowling Green, Ohio 43403

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Ronald N. Marso
Fred L. Pigge
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

A paper presented at the annual meeting of the
Mid-Western Educational Research Association
Chicago, Illinois
October 19-21, 1989

Running Head: CLASSROOM TESTING STAFF DEVELOPMENT

Abstract

This paper presents several staff development implications for elementary and secondary teachers which were gleaned from 900 supervisors', principals', and teachers' assessments of classroom teachers' testing needs, resources, practices, and proficiencies. These assessments included direct analyses of teacher-made tests as well as perceptual assessments of teachers' testing needs and proficiencies. Generally, it was agreed that teachers' testing proficiencies were inadequate to meet classroom needs and analyses showed that typical teachers gave 50 or more formal teacher-made tests per year and wrote most of their own test questions; matching exercises on teacher-made tests were particularly error prone; most teacher-made tests except in the math or science content areas functioned primarily at the knowledge level; administrators' and teachers' perceptual assessments of teachers' testing skills were negatively correlated with the results of the direct analyses of teacher testing skills as displayed on their teacher-made tests; standardized tests were reported to be less needed in the secondary than in the elementary grades; and schools had very limited resources available to support teachers' testing responsibilities including basic typing and duplication services.

BEST COPY AVAILABLE

Staff Development Implications From A State-Wide Assessment of Classroom Teachers' Testing Skills and Practices

Despite the rather extensive literature advising classroom teachers regarding the development and use of teacher-made tests, little is really known about teachers' tests or their day-to-day classroom testing practices; only recently have empirical investigations of classroom teachers' testing practices, testing attitudes, and test construction proficiencies appeared in the literature. Coffman (1971) in the early 70's stated that relatively little research had been conducted on teacher-made tests and related testing practices in the public schools, and a decade later Dwyer (1982) still described advice to teachers about teacher-made tests and related testing practices to be based upon a consensus of professional judgments rather than upon empirical investigations. In one of the empirical reports on classroom teacher testing, Fleming and Chambers (1983) described teacher testing as a "window" to on the classroom with teacher-made tests occupying considerable amounts of classroom time but being a little researched part of the drama in classrooms.

Several of the recent empirical investigations of classroom testing practices have added to our knowledge about teacher testing practices and also suggest that public school teachers have limited testing proficiencies and may not put into practice procedures typically taught during teacher training. For example, Gullickson (1984) reported that teachers schedule frequent tests and are supportive of testing, but they desire more assistance in better meeting their testing responsibilities. Stiggins and Bridgeford (1985) found that testing procedures varied by both grade level and subject area and also reported that their sample of teachers expressed concerns about improving their tests and about having insufficient time to do so. And, Gullickson and Ellwein (1985) reported that in a relative sense very few teachers regularly use post-hoc statistical analyses after administering their teacher-made tests contrary to such emphasis in teacher-preparation tests and measurement courses.

Other research conducted in the public schools suggest that certain testing practices do have an impact upon what is happening in classrooms. For example, some research findings suggest that students prepare differently for different test item types (D'Ydewalle, Swerts, & DeCorte, 1983; Kulhooy, Dyer, & Silver, 1975), that some types but not others of test feedback enhances learning (Hanna, 1976; Stewart & White, 1976; Wexley & Thornton, 1972), that students have preference for certain test item types (Shaha, 1984), that frequent scheduling of tests designed to enhance learning actually does so (Peckham & Roe, 1977), and that class time spent testing is more efficient in facilitating learning than is time spent reviewing content (Nungester & Duchastel, 1982).

A few investigations of classroom teachers' testing proficiencies or practices have been based upon direct analyses of teacher-made tests. Fleming and Chambers (1983) conducted an assessment of 342 teacher-made tests and concluded that short response (including fill-in-the-blank) items and matching exercises were the most frequently used item types with essay item types infrequently used regardless of grade level; that most test items measured only at the knowledge level (69% to 94% at varied grade levels); that those items

measuring beyond the knowledge level were found almost exclusively on just the math and science tests; and that test format type construction errors were very common (lack of consecutive item numbering, handwritten and illegible text, and spelling, grammatical, and punctuation errors). Billeh (1974) and Black (1980) also reported studies based upon the direct analyses of teacher-made tests, but their investigations were restricted to an assessment of the cognitive functioning levels of test items found just on science tests. They found that test item cognitive functioning levels varied by science subject with biology and chemistry tests having the highest levels of knowledge level items (66% to 94%) and with physics tests containing the most desirable range of cognitive functioning items.

Purpose

The purpose of this paper is to present selected findings which suggest implications for staff development from an investigation of teacher-made testing needs, proficiencies, practices, and resources in Ohio based upon direct analyses of actual teacher-made tests as well as upon perceptual assessments by teachers, principals, and supervisors. As indicated, just selected findings are presented in this paper as detailed discussions of the findings from the broader study are reported elsewhere (Marso & Pigge, 1989; Marso & Pigge, 1988a; Marso & Pigge, 1988b); thus the goal herein is to present and discuss those findings from the larger study of teacher testing which relate to suggestions for staff support or staff development in the public schools.

The following types of questions are illustrative of the focus of this paper: a) Are teachers' testing proficiencies thought to be adequate in meeting their classroom testing needs? b) Are resources in the schools sufficient to support teacher testing responsibilities? c) What are the most common types of test construction errors found on teacher-made tests? and d) What are typical classroom teachers' testing practices, such as: How frequently do classroom teachers schedule formal exams?

What item types are most frequently used in constructing teacher-made tests? What statistical analyses are used on test scores obtained from teacher-made tests? Do teachers construct their own test items?

Methods and Procedures

The Subjects

One group of subjects for this study consisted of 800 Ohio public school supervisors and principals whose names were randomly selected from the state directory of schools. The type of school system (city, exempted village, and county local), and grade level assignment (elementary, middle, and secondary) classifications used in the directory were also used as strata in the random selection of the administrators. After two follow-up mailed contacts to nonrespondents, 586 (73%) of these administrators completed the survey assessment of which 229 were classified as teacher supervisors, 313 principals, and 44 as individuals in related teacher supervisory roles (coordinators of curriculum or instruction, etc.).

The second group of subjects were selected by "matching" the social security numbers of Bowling Green State University graduates during the years of 1975 through 1985 with the social security numbers of full-time teachers certified by the Ohio State Department of Education for the 1985-86 school year. This procedure resulted in the identification of 600 teachers from whom usable survey assessment responses were obtained from 326 (54%). Only teachers with regular classroom assignments were selected for the study (specialized area teachers were excluded, e.g., art, music, special education, etc.). Each teacher, regardless of teaching field, had been required by BGSU to take a tests and measurements course during teacher training.

Instrumentation

The assessment instrument was comprised of 45 testing competency descriptions or items placed within four categories: working with teacher-made tests, using teacher-made test scores, working with purchased tests and scores, and working with competency or mastery testing programs. All respondents were directed to respond to the competency-mastery testing section only if their schools were involved in such programs. Both the administrator and teacher forms of the assessment instrument contained these four sections of items presented in identical format. Each of the 45 testing competency items was responded to via two five-point Likert-type scales marked from high (5) to low (1) with the two scale headings for the administrators' form being: "need of this competency to be a successful teacher in your school" and "average proficiency of your new teachers in this competency." The two-scale headings for the teachers' form were: "to be successful in your job, what is your need for this competency" and "an estimate of your classroom proficiency in this area."

In addition to the 45 testing competencies presented on the assessment instrument, both the administrators and the teachers were asked to report the extent of the availability of 12 resources or policy guidelines to support teachers' testing responsibilities in their schools. Both the administrators and teachers were also asked to assess, in a comparative sense, the level of their training (teachers' form) or typical beginning teachers' proficiencies (administrators' form) in testing and evaluation via three Likert-type scale items relative to: a) teachers' knowledge of their subject area, b) teachers' other professional education competencies, such as planning lessons, handling discipline, etc., and c) teachers' overall competencies as educators. The teacher form also contained one additional section asking the teachers to report on seven of their testing preferences and practices such as how frequently they scheduled formal teacher-made tests and what types of test items they most commonly used in developing their classroom tests.

Sample of Teacher-Made Tests

All teachers were also asked to enclose a copy of their most recently developed formal teacher-made test (not a quiz or a test from spelling or math class unless they were a math major) when returning the completed assessment instrument. This resulted in the collection of 175 (54%) teacher-made tests. These tests, regardless of grade level, when classified by subject area

consisted of 30 tests pertaining to history/social studies, 36 science, 29 business education, 32 mathematics, 28 English, and 20 tests pertaining to nine other content specializations with insufficient numbers to be included as distinct subject area categories.

The sample of 175 teacher-made tests included a total of 6504 test items and 455 item exercises (a group of items of a similar item type). The test items within the sample of tests were classified independently by two judges using Bloom's taxonomy of six cognitive demand levels (knowledge, comprehension, application, analysis, synthesis, and evaluation). If the judges differed in their classification of an item or exercise, the item or exercise was reexamined until a consensus was reached between the two judges.

Each test and each test exercise was also examined for format and item construction errors. Item and test format construction error criteria were selected from a review of several tests and measurements textbooks designed for preservice education courses. A total of eight item type classifications (completion, essay, multiple-choice, etc.), 10 item format construction error criteria (does the test have complete directions? are item types grouped together? are the items numbered consecutively? etc.), and 66 item construction error criteria (incomplete stems, implausible alternates, specific determiners, etc.) were identified from these procedures and used in the assessment of the sample of teacher-made tests (see Tables 1 and 2 for a listing of these types and criteria).

An item construction error, if present, was recorded once per item exercise rather than each time that particular error type may have occurred within an item exercise. In other words, whether or not a construction error appeared just on one item or on several items within the same item exercise, a tally of '1' was recorded for that particular error. Similarly, each test format error was recorded only once per entire test regardless of whether or not that error occurred more than once on a particular test. It was thought that this procedure would provide a more stable base of comparison across tests which varied in their number of test items and item exercises.

Implications, Findings, and Rationale

Implication #1. Building principals and supervisors need to increase the availability of testing resources, policy or guideline statements, and support services to better assist teachers in meeting their testing responsibilities. This implication is based on the following findings.

Related Findings

1.a Just 50% of the teachers reported that typing and duplication services were consistently available to support their testing responsibilities.

1.b Just 7 to 15% of the teachers reported the availability of grade assignment, grade frequency, or term grade calculation guidelines in their schools.

1.c Just 15 to 26% of the teachers reported the availability of computer assistance services such as test scoring, statistical analyses, or test generation software.

1.d Beginning teachers in particular were less likely to possess textbook instructor manuals with test items and related resources to support their testing responsibilities.

1.e Testing demands much teacher time and effort, the average number of tests given per teacher was 54.1 per academic year and approximately one-half of the teachers reported writing 75% of the items comprising their tests (37% reported writing nearly all their items).

Rationale

Teachers have many demands on their time; the availability of appropriate testing resources, guidelines, and support services should have a significant positive impact on teacher effectiveness. Administrators should (a) secure the assistance of student interns or parent volunteers or seek financial resources to employ assistance in typing and duplication of tests, (b) coordinate the development of grading guidelines, and (c) make certain that instructor manuals are available, especially for new teachers. For example, it is commonly known that the assignment of term grades to pupils demands considerable teacher effort which frequently results in inequities and feelings of frustration and also often leads to conflict between teachers and students, between teachers and parents, and between teachers themselves. The development of departmental or building pupil term grade assignment guidelines could well save a considerable amount of teachers' time, frustration, and conflict.

Implication #2. Teachers' inservice training on skills related to teacher-made test construction and use is needed and desired. (Many universities do not offer a separate course in tests and measurements in their teacher training program.) Relatedly similar training should be provided for principals and supervisors so that they can better supervise and assist teachers in constructing and using teacher-made tests. This implication is based upon the following findings.

Related Findings

2.a The supervisors and principals rated the need for classroom testing competencies higher than they rated typical beginning teachers' testing proficiencies. (Other evidence collected revealed that these competencies did not improve with teaching experience.)

2.b The teachers reported a desire to improve their teacher-made tests.

2.c Comparatively, the teachers, principals, and supervisors rated teachers' testing competencies lower than they rated teachers' other professional competencies and proficiencies.

2.d Both the teachers and the administrators reported a high need for teachers to possess skills in the use of teacher-made and standardized tests in day-to-day instruction to assure teachers' success in the classroom.

2.e Neither the administrators' nor the teachers' ratings of teachers' test item construction skills were supported by the results of the direct analyses of teachers' item writing skills as displayed on their teacher-made tests (negative correlations [in the -.50s to -.70s range] were found between the direct analyses of teachers' item writing skills and administrators' and teachers' perceived ratings of these skills).

Rationale

The teachers, principals, and supervisors appeared to be in consensus regarding their perceptions that testing skills are among teachers' weaker professional proficiencies. This perception was confirmed to some extent by the direct assessment of the sample of teacher-made tests. The findings also suggested that teachers, principals, and supervisors may need to be trained to better identify test construction errors so that efficient steps can be taken to alleviate these errors.

Implication #3. When assessing teachers' inservice training needs (at least in the area of testing but probably in other areas as well), inservice trainers should not combine teachers', principals', and supervisors' ratings of teachers' inservice needs as each of these groups tends to rate competencies at different need intensity levels than do the other groups; the combined averages conceal these differences. Rather, the trainers should probably address those proficiencies rated lowest within each individual group's ratings. This implication is based upon the following findings.

Related Findings

3.a The teachers', principals', and supervisors' ratings indicated relative rank-order agreement among one another on both the classroom need for and the teachers' proficiency in most testing competencies (although the magnitudes of their ratings differ).

3.b The relative agreement (high rank-order correlations) between principals' and supervisors' ratings of both teachers' testing needs and teachers' testing proficiencies is very high; less agreement (moderately high correlations) is evident between administrators' and teachers' ratings of teachers' testing needs and proficiencies.

3.c The teachers' ratings of teachers' testing proficiencies were highest followed by principals' ratings, and then by supervisors' ratings.

Rationale

Teachers, principals, and supervisors whether due to differences in their individual nature, due to their differing insight into daily classroom operations, or due to other unknown factors rate teachers' testing proficiencies at different intensity levels. It would appear that administrators' ratings of

teachers' testing proficiencies should be solicited for a second perspective of teachers' performance in planning inservice training, but these ratings should be interpreted separately from teachers' ratings.

Implication #4. Testing specialists and/or practicing educators need to re-examine the role and value of test construction statistics for classroom teachers. Practicing educators (teachers, supervisors, and principals) reported little need for statistical analyses (calculation of means, standard deviations, reliability, etc.) of teacher-made test results which is in apparent contrast to the emphasis placed upon these skills by testing specialists and writers of preservice tests and measurements textbooks. This implication is based upon the following findings.

Related Findings

4.a The teachers rated the need for the statistical analyses of teacher-made test results the lowest among the presented 45 testing competencies.

4.b The school administrators (principals and supervisors) rated the need for the statistical analyses of teacher-made test results the lowest among the 45 competencies.

4.c The school administrators (principals and supervisors) rated teachers' proficiency in statistical analyses of teacher-made test results the lowest among the 45 testing competencies.

4.d Most of the teachers reported that they never or rarely calculate means or standard deviations (80%), never or rarely complete item analyses (54%), and never or rarely estimate test reliability (60%) for their teacher-made tests.

Rationale

Testing specialists have not been successful in convincing practicing educators of the value of the statistical analyses of teacher-made test results, or they, themselves, are mistaken about the value of these procedures for practicing educators.

Implication #5. Less emphasis needs to be placed on standardized testing topics in preservice and inservice training for secondary level teachers as compared with elementary level teachers. Principals', supervisors', and teachers see less classroom need for teachers' proficiencies in competencies related to working with standardized tests in the upper grade levels as compared to the elementary grades. This implication is based upon the following findings.

Related Findings

5.a The principals and supervisors perceived less need for standardized testing competencies in the middle and the secondary grades as compared to the elementary grades.

5.b The principals' and supervisors' rated teachers' proficiencies in standardized testing competencies somewhat lower in the middle and secondary as compared to the elementary grades.

5.c The teachers and administrators rated both classroom need and teachers' proficiency in working with standardized tests somewhat lower than the comparable ratings for teacher-made tests.

Rationale

Principals', supervisors', and teachers' ratings of classroom needs and teachers' proficiencies in working with standardized tests may simply reflect typical school practices. Namely, teacher-made tests are given much more frequently than standardized achievement tests, and standardized achievement tests tend to be more frequently given and usually are more integrated into the instructional processes in the lower grades as compared to the secondary grades. Implication #6. Inservice instruction of teachers in tests and measurements skills need not vary significantly in content for teachers with more or less teaching experience (at least between 1 to 10 years) or for teachers employed in different types of school settings (urban, rural, and suburban) although inservice trainers need to be aware that the availability of testing resources will likely vary by type of school setting. This implication is based upon the following findings.

Related Findings

6.a Neither the teachers' testing proficiencies nor classroom testing needs were found to differ between rural, urban, and suburban schools settings.

6.b The availability of testing resources were found to vary by school setting with more resources being available in suburban schools and fewer resources available in urban schools. Thus, if success of the pre-service or inservice training being provided depends upon the availability of school testing resources adjustments may be necessary.

6.c Neither the teachers' ratings of classroom testing need nor of their proficiency in the 45 testing competencies varied with the years of their teaching experiences (1 to 10 years). Thus, it would appear that experience in teaching, in itself, does not result in improved teacher-made tests or in a different set of needs.

Rationale

The nature of testing competencies appears to be such that teachers' needs and proficiencies do not vary by type of school setting or through years of teaching experience. It is not known if inservice training related to testing skills had been provided to this sample of teachers during their years of teaching experience or, if so, whether such training had been ineffective. Other research does indicate that testing skills tend not to increase through practice without feedback. This would suggest that if feedback for teachers' test construction performance had been provided by supervisors or principals,

then teachers' test construction skills might have increased with teaching and testing experience.

Implication #7. The content emphasis for teacher inservice training in classroom test construction should vary by teachers' grade level assignments and for subject area specializations, at least in terms of focus on writing different item types. This implication is based upon the following findings.

Related Findings

7.a The secondary teachers were more likely than elementary teachers to have used problem and essay type items; the elementary teachers were more likely to have used completion and multiple-choice items than the secondary teachers.

7.b The secondary teachers reported that they more frequently used statistical analyses with their test results and that they wrote more of their own test items than did the elementary grade teachers.

7.c Teachers in most subject areas used a variety of item types with the exceptions that English and social studies teachers were unlikely to have used problem item types and that math and lower grade teachers were not likely to have used completion (fill in the blank) or essay items.

Rationale

Some test item types lend themselves to particular subject area content and not to others, some item types are inappropriate for the very low grade levels, and in some cases precedent or past practices in a field of specialization may tend to limit teachers' use of a variety of item types.

Implication #8. Teachers' (and public school administrators') inservice training in test construction should include the development of skills in the identification of common types of item construction errors and in selecting good items from already constructed pools of test items. This implication is based upon the following findings.

Related Findings

8.a The negative correlations found between teachers' (and administrators') perceived rating of item writing proficiencies and actual teachers item writing proficiencies as displayed on their teacher-made tests (magnitudes of -.50 to -.70) suggested that teachers (and administrators) may not be able to effectively identify test construction errors in written item exercises. Secondly, teachers with several years of teaching (up to 10 years) made the same types and about the same frequencies of errors as did teachers with very few years of teaching experience.

8.b The analyses of the actual teacher-made tests indicated that most of the high frequency item construction errors are given major attention in the tests and measurements textbooks (and presumably in course instruction)

used in preservice education. This suggests that preservice instruction in this area is not very effective, that teachers do not know how to identify or alleviate these errors, or that time restraints, desire, and/or other factors prevent teachers from eliminating these types of errors in their tests.

8.c Approximately one-half of the classroom teachers indicated that they select 25% or more of the test items used in their formal exams from other sources to supplement the items that they, themselves, write.

Rationale

Teachers (and administrators) need skills in selecting items from old tests, textbook instructor manuals, and other sources as well as in identifying errors in items which they have constructed themselves in order to improve their teacher-made tests. If teachers and their supervisors cannot recognize errors in test items, then teachers will not feel the need to improve their test items.

Implication #9. Teachers' inservice training in test construction and use should include the development of skills in classifying existing test items by cognitive functioning level and in the construction of test items that measure beyond the knowledge level (e.g., Bloom's six cognitive levels of functioning: knowledge, comprehension, application, analysis, synthesis, and evaluation). This implication is based upon the following findings.

Related Findings

9.a Of the 6,529 items examined on the teacher-made tests, 72% of the items functioned at the knowledge (simple recall) cognitive level (11% measured at comprehension, 15% application, and 1% higher than application.).

9.b With the exception of the math and science tests, nearly all of the sampled teacher-made tests functioned almost completely (95% or more of the items) at the knowledge level (over one-half of all items measuring beyond the knowledge level were found on the math and science tests).

9.c The social studies tests as a group functioned almost completely at the knowledge level (98% of the total group of these items were judged to be functioning at the knowledge level).

9.d Both the teachers and administrators reported a very high need for teachers to write test items that measure pupils' higher-order thinking processes.

Rationale

Testing specialists are very concerned with the cognitive functioning levels of teacher-made tests for at least two reasons. One, as the attainment of many educational goals requires performance beyond the knowledge level, it is assumed that a test measuring at only the knowledge level cannot be a valid index of progress in attaining these types of objectives. Secondly, as students tend to study what is tested, tests measuring exclusively at the knowledge level

then may be encouraging students to study the less significant elements of class content. In other words, tests measuring just at the knowledge level may be impeding the learning-instructional process.

Implication #10. Inservice training of teachers in test construction and use should include skill development in writing most item types as most teachers tend to use a relatively wide variety of item types. This implication is based upon the following findings.

Related Findings

10.a Only the essay (not found in lower grades) and problem (limited primarily to math/science content) item types were found to have a severely restricted content or grade level use when the teacher-made tests were examined.

10.b Most of the teachers used a variety of item types; an average 2.6 item types were found on the teacher-made tests.

10.c Relative to frequency of item type used on the sample of teacher-made tests, 51% contained short response items, 45% contained matching exercises, 39% contained true-false, 37% contained multiple-choice, 31% contained problems, 27% contained completion, 17% contained interpretive exercises, and 13% contained essay type items.

10.d These observations were noted in terms of percentage of the use of each item type relative to the total number of items found on the teacher-made tests: 20% were multiple-choice, 19% were matching, 17% were short responses, 14% were true-false, 14% were problems, 6% were completion, 6% were interpretive exercises, and 1% were essay type items.

10.e The most commonly used item type on the teacher-made tests in terms of both the frequency of appearance and the numbers of items written was the completion-short response form of question.

Rationale

Test specialists encourage the use of a relatively wide variety of item types for the following reasons: a) students tend to study somewhat differently for different item types (the use of a variety of item types encourages more diverse study habits), b) selection of item type limits cognitive functioning levels of the test (i.e., the short-response and completion items seldom measure beyond knowledge level which may in part account for the preponderance of knowledge functioning items found on teacher-made tests), c) some item types better "fit" certain subject-content, and d) some students prefer some item types and some students prefer other types.

Implication #11. Most inservice training related to test item construction skills need not be extensive nor particularly technical to be successful. This implication is based upon the following findings.

Related Findings

1. Inservice training in teacher test construction need not be presented just by test specialists as many types of errors found on the teacher-made tests were found to be of the nontechnical variety.

2. The item construction errors typically addressed in introductory tests and measurements textbooks account for nearly all errors found on teacher-made tests.

3. Inservice training related to item writing and test format skills if focused just on the two to four most frequently occurring construction errors for each item type (with the exception of the highly error prone matching exercises) would address the majority of errors found on teacher-made tests. The most frequently identified types of errors can be ascertained by reviewing Tables 1 and 2.

References

Black, T. R. (1980). An analysis of levels of thinking in Nigerian science teachers' examinations. Journal of Research in Science Teaching, 17, 301-306.

Billeh, V. Y. (1974). An analysis of teacher-made test items in light of the taxonomic objectives of education. Science Education, 58, 313-319.

Coffman, W. E. (1971). Essay examinations. In R. L. Thorndike (Ed.), Educational measurement (2nd ed., pp. 271-302). Washington, D.C.: American Council on Education.

Dwyer, C. A. (1982). Achievement testing. In h. E. Mitzel (Ed.), Encyclopedia of educational research (4th ed., Vol. 1, pp. 13-22). New York: The Free Press.

D'Ydewalle, G., Swerts, A., & DeCorte, E. (1983). Study time and test performance as a function of test expections. Contemporary Educational Psychology, 8, 55-67.

Fleming, M. & Chambers, B. (1983). Teacher-made tests: Windows on the classroom. New directions for testing and measurement, 19, 29-38.

Gullickson, A. R. (1984). Teacher perspectives of their instructional use of tests. Journal of Educational Research, 77, 244-248.

Gullickson, A. R., & Ellwein, M. C. (1985). Post hoc analysis of teacher-made tests: The goodness-of-fit between prescription and practice. Educational Measurement: Issues and Practice, Spring, 15-18.

Hanna, G. S. (1976). Effects of total and partial feedback in multiple-choice testing upon learning. Journal of Educational Research, 69, 202-205.

Kulhooy, R. W., Dyer, J. W., & Silver, L. (1975). The effects of notetaking and test expectancy on the learning of text material. Journal of Educational Research, 68, 363-365.

Marso, R. N., & Pigge, F. L. (1989). The status of classroom teachers' test construction proficiencies: Assessments by teachers, principals, and supervisors validated by analysis of actual teacher-made tests. ERIC Document TM 013142.

Marso, R. N., & Pigge, F. L. (1988a). Teacher-made tests and testing: Classroom resources, guidelines, and practices. ERIC Document: ED TM 0116066.

Marso, R. N., & Pigge, F. L. (1988b). An analysis of teacher-made tests: Testing practices, cogntive demands, and item construction errors. ERIC Document: ED TM 012323.

Nungester, R. J., & Duchastel, P. C. (1982). Testing versus review: Effects on retention. Journal of Educational Psychology, 74, 18-22.

Peckham, P. D., & Roe, M. D. (1977). The effects of frequent testing. Journal of Research and Development in Education, 10, 40-50.

Shaha, S. (1984). Marching-tests: Reduced anxiety and increased test effectiveness. Educational and Psychological Measurement, 44, 869-881.

Stewart, L. G., & White, M. A. (1976). Teacher comments, letter grades, and student performance: What do we really know? Journal of Educational Psychology, 68, 488-500.

Stiggins, R. J., & Bridgeford, N. J. (1985). The ecology of classroom assessment. Journal of Educational Measurement, 22, 271-286.

Wexley, K. N., & Thornton, C. L. (1972). Effect of verbal feedback of test results upon learning. Journal of Educational Research, 66, 119-121.

8/4

Table 1
Test Construction Error Summaries

	No. Items Reviewed	% Total Items Reviewed	No. of Exercises	No. Errors Present*	Mean Errors Per Exercise
A. Item Type					
1. Matching	1261	19	78	496	6.4
2. Completion	549	8	48	106	2.2
3. Essay	64	1	22	34	1.5
4. True/False	935	14	69	71	1.0
5. Multiple-Choice	1317	20	65	53	.8
6. Short Response	1093	17	89	61	.7
7. Problems	896	14	54	26	.5
8. Interpretive Exercise	362	6	30	6	.2
9. Unclassified	52	1	6	-	-
Subtotals	6529	99	455	853	1.9

B. Test Format Errors	No. Tests** Where Errors Present	% of Total
1. Absence of directions	82	29
2. Answering procedures unclear	61	22
3. Items not consecutively numbered	47	17
4. Adequate margins	22	8
5. Answer space provided	21	7
6. Space between items	12	4
7. Nonindependent items	11	4
8. Different weighting of objective items	8	3
9. Items arrange most to least time demanding	7	2
10. Similar item types not grouped together	6	2
	281	100

*Each specific item type construction error was tallied only once if present in an exercise (i.e., an error may have occurred several times or once in an exercise but in either case only a single tally was used so that tests and exercises could be compared regardless of the number of individual items appearing in a test or exercise).

Table 2
Frequency and Nature of Item Construction Errors Found for Each Type of Item Exercise

Construction Error	N	%*
a. Completion Item Type		
Not complete interrogative sentence	32	30
Blanks in statements	31	29
Textbook statements with words left out	18	17
More than single blank in statement	12	11
Question allows more than single answer	6	6
Blank number clue	4	1
Blank length clue	1	1
Requests trivia versus significant idea	1	1
Unstated degree of precision	1	1
Lengthy, unnecessary words or phrases	0	0
	106	100
b. True-False		
Required to write response, time waste	20	28
Statements contain more than single idea	16	23
Negative statements used	15	21
Presence of specific determiner	8	11
Statement not question, give away item	6	8
Needless phrases present, too lengthy	4	6
Imprecise statement, not always true or false	1	2
Presence of length clue	1	1
Opinion not attributed to source	0	0
	71	100
c. Essay Exercises		
Response expectations unclear, not labeled, etc.	14	41
Scoring points not realistically limited	7	21
Optional questions provided	5	15
Restricted question not provided	3	9
Ambiguous words used	2	6
Opinion or feelings requested	2	6
Question limited to simple listing response	1	2
	34	100
d. Problem Exercises		
Items not sample understanding concepts, only calculations	20	77
Not range of easy to difficult problems	3	12
Degree of accuracy not requested	2	8
Nonindependent items	1	4
Use of objective items when calculation preferable	0	0
	26	100

(table continues)

Construction Error	N	%*
e. Matching Item Type		
Columns not titled	71	14
Not use one, more than once, or not all not in directions to prevent elimination	69	14
Response column not ordered	60	12
Directions not specify basis for match	55	11
Answering procedure not specified	52	10
Elimination due to equal numbers	46	9
Column(s) exceed 10 items	39	8
Materials not homogeneous	38	8
Premise not to left side	37	7
Numbers not to left and letters to right	13	3
Exercise not contained on single page	7	2
Requires responses to be written out	6	1
Insufficient information in premises	3	1
	496	100
g. Interpretive Exercises		
Objective response form not used	6	100
Can be answered without data presented	0	0
Errors present in response items	0	0
Data presented unclear	0	0
	6	100

Construction Error	N	%*
f. Multiple Choice		
Alternates not in column(s) or rows	21	40
Incomplete stems	12	23
Negative words not emphasized or avoided	9	17
"All or none above" not appropriately used	5	9
Needless repetition in alternates	2	4
Presence of specific determiners in alternates	2	4
Verbal associations between alternate and stem	1	1
Alternates overlap	1	1
Needless phrases used	0	0
Grammatical clues	0	0
Distractors implausible	0	0
Length clues	0	0
a and c, but not b, etc. used	0	0
	53	100
h. Short Response		
Item requires only listing	51	84
Response expectations ambiguous, not specified	7	11
Unrealistically high scoring values assigned	3	5
	61	100

*Each specific item type construction error was tallied only once if present in an exercise (i.e., an error may have occurred several times or once in an exercise but in either case only a single tally was used so that tests and exercises could be compared regardless of the number of individual items appearing in a test or exercise), the percentage refers to percent of this error type to all errors found on all exercises of this type.

8/4